Research Report

Interpretations of Reliability

Shelby J. Haberman

Research & Development

December 2005
RR-05-29

Interpretations of Reliability

Shelby J. Haberman
ETS, Princeton, NJ

December 2005

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Some probabilistic illustrations of the reliability coefficient are provided to assist in interpretation of this measure. All explanations are derived under the assumption that the joint distribution of examinee scores from two parallel tests is well approximated by a bivariate normal distribution.

Key words: Concordance, cut points, interval estimation

Acknowledgements

The author would like to thank Tim Davey and Skip Livingston for helpful comments.

A number of probabilistic interpretations of reliability coefficients are readily available given concepts of simple random sampling from a very large population and given a bivariate normal approximation for the joint distribution of test scores from two parallel test forms. In this note, three such interpretations are provided. The first considers the probability that, if two examinees are selected at random and scores on Forms 1 and 2 are recorded, then the same examinee obtains the higher score on both forms. The second considers the probability that an examinee who has exceeded a cut-point score on Form 1 also exceeds the cut point on Form 2. The third interpretation considers the interval width based on Form 1 that suffices to include the examinee score on Form 2 with a given probability.

In each example, it is an elementary matter to provide a table that indicates, for a given reliability, what is the corresponding probability or interval width. Thus the tables may have some potential for use in educating relatively less technically oriented audiences about the meaning of reliability. The mathematical formulas for computation of table entries are not difficult to derive, but they are not themselves readily understood by audiences that are not mathematically sophisticated.

1. Constant Order

Let two examinees numbered 1 and 2 be obtained by simple random sampling without replacement from a very large population of examinees. Let examinee k, k equals 1 or 2, have score X_{jk} on form j. Given the assumption of simple random sampling, the joint distribution of scores X_{11} and X_{21} for Examinee 1 is the same as the joint distribution of the scores X_{12} and X_{22} for Examinee 2. Given the assumption that the population is very large, scores for Examinee 1 can be regarded as independent of scores for Examinee 2. For simplicity, assume that the joint distribution of X_{1k} and X_{2k} is well approximated by a bivariate normal distribution. The assumption that the forms are parallel implies that X_{1k} and X_{2k} have common mean μ and common standard deviation σ for each examinee k. Let the reliability coefficient be ρ^2, so that ρ^2 is the correlation of X_{1k} and X_{2k} for each examinee k.

Consider the concordance probability C that the same examinee has the higher score on both examinations. Under the bivariate normal approximation, this probability, which is encountered in the study of Kendall's τ (Kruskal, 1958), is equal to $0.5 + \pi^{-1} \sin^{-1}(\rho^2)$. Table 1 provides a table of C and ρ^2.

Table 1.
Relationship of Reliability and Concordance

Reliability	Concordance (form to form)	Concordance (form to true)
0.0	0.50	0.50
0.1	0.53	0.60
0.2	0.56	0.65
0.3	0.60	0.68
0.4	0.63	0.72
0.5	0.67	0.75
0.6	0.70	0.78
0.7	0.75	0.82
0.8	0.80	0.85
0.9	0.86	0.90

One simple criterion based on the table is the point at which the concordance probability has progressed half way from its minimum value of 0.5 to its maximum value of 1. This point is reached at $\rho^2 = 1/2^{1/2} = 0.71$.

Interestingly, the choice of $\rho^2 = 1/2^{1/2}$ arises from a very different case. For examinee k, consider the best predictor of score X_{2k} on Form 2 from score X_{1k} on Form 1. Under the bivariate normal approximation, the mean-squared error of this predictor is $\sigma^2(1-\rho^4)$. Without knowledge of X_{1k}, σ^2 is the best mean-squared error achievable by prediction of X_{2k} by a constant c. If $\rho^2 = 1/2^{1/2}$, then the mean-squared error $\sigma^2/2$ from optimal prediction of X_{2k} from X_{1k} is half the mean-squared error σ^2 from optimal prediction of X_{2k} by a constant.

Results in this section are more conservative than those obtained from an interpretation based on a comparison of rankings based on true scores and on observed scores. For each examinee k, there exists a variable T_k, the true score of examinee k, with mean μ such that $X_{jk} = T_k + e_{jk}$ for form j and such that T_k, e_{jk}, and e_{jk} are uncorrelated for each form j and examinee k. The standard deviation of T_k is $\sigma(1-\rho^2)^{1/2}$, the standard deviation of e_{jk} is $\sigma\rho$, and the correlation of X_{jk} and T_k is ρ. Under the normal approximation, the variables T_1, T_2, e_{11}, e_{21}, e_{12}, and e_{22} are mutually independent (Lord & Novick, 1968, chap. 3).

Under the normal approximation, the concordance probability C_T that the same examinee has the higher score on form k and has the higher true score is equal to $0.5 + \pi^{-1}\sin^{-1}(\rho)$. Results are listed in Table 1. In this case, the concordance probability of 0.75 is attained for a reliability

of only 0.5. Nonetheless, very high concordance probabilities still require quite high reliability, as is evident from the probability of 0.90 for the reliability of 0.90.

2. Cut Points

As an alternative interpretation, consider use of cut points. Suppose that a score is acceptable if it falls above the $100p$th percentile for some p greater than 0 and less than 1. Alternatively, a cut point z may have been selected, and p is the probability that a randomly selected examinee scores below z. Consider the conditional probability that an examinee receives an acceptable score on Form 2 given that the examinee receives an acceptable score on Form 1. Let ϕ denote the standard normal density, let Φ denote the standard normal distribution function, and let $q = \Phi^{-1}(p)$ be the standard normal percentile that corresponds to $100p$. Then elementary arguments may be used to show that the normal approximation yields the joint probability

$$J = \int [1 - \Phi([q - \rho y]/(1 - \rho^2)^{1/2})]^2 \phi(y) dy$$

that a randomly selected examinee receives an acceptable score on both Form 1 and Form 2, and $J/(1 - p)$ is the corresponding conditional probability that the examinee receives an acceptable score on Form 2 given that an acceptable score is received on Form 1. Results are provided in Table 2.

The table suggests that exceeding a cut point once does not provide much assurance of exceeding a cut point again even with rather high reliability. Reliability does matter, for results for $\rho^2 = 0.9$ are considerably better than for $\rho^2 = 0.8$. The greatest challenge is for high cut points. Even for a reliability of 0.9, for $100p = 80$, the conditional probability is only 0.75 that the cut point on Form 2 is exceeded given that the cut point on Form 1 is exceeded.

More favorable results are obtained if one considers the probability that the score X_{jk} exceeds the cut point given that the true score T_k exceeds the cut point. In this instance, the conditional probability sought is

$$(1 - p)^{-1} \int_q^\infty [1 - \Phi([q - \rho y]/(1 - \rho^2)^{1/2})] \phi(y) dy.$$

For some results, see Table 2. Note that high cut points still present challenges, as is evident from the case of $100p = 80$ and $\rho^2 = 0.9$.

Table 2.
Probabilities of Exceeding Cut Points

Percentile	Reliability	Joint probability (two forms)	Conditional probability (Form 2 given Form 1)	Conditional probability (form given true)
20	0.0	0.64	0.80	0.80
40	0.0	0.36	0.60	0.60
60	0.0	0.16	0.40	0.40
80	0.0	0.04	0.20	0.20
20	0.2	0.66	0.82	0.85
40	0.2	0.39	0.65	0.72
60	0.2	0.19	0.48	0.58
80	0.2	0.06	0.28	0.41
20	0.4	0.68	0.85	0.88
40	0.4	0.42	0.70	0.77
60	0.4	0.22	0.56	0.66
80	0.4	0.08	0.38	0.52
20	0.6	0.70	0.87	0.91
40	0.6	0.46	0.76	0.82
60	0.6	0.26	0.64	0.74
80	0.6	0.10	0.50	0.62
20	0.8	0.73	0.91	0.94
40	0.8	0.50	0.83	0.88
60	0.8	0.30	0.75	0.82
80	0.8	0.13	0.65	0.74
20	0.9	0.75	0.94	0.95
40	0.9	0.53	0.88	0.92
60	0.9	0.33	0.83	0.88
80	0.9	0.15	0.75	0.82

3. Intervals

Consider use of the score X_{1k} on Form 1 to provide an interval that contains the score X_{2k} on Form 2 with a given probability p. If $z = \Phi^{-1}(1 - p/2)$, then a suitable interval based on the normal approximation is centered at $\mu + \rho^2(X_1 - \mu)$ and has width $2z\sigma(1 - \rho^4)^{1/2}$. Table 3 provides intervals for $p = 0.05$, so that the coverage probability is 0.95, and for $\sigma = 100$, a value relatively close to that encountered with the SAT® I math or verbal examination. For narrower intervals, the case of $p = 0.5$ is also considered, so that the coverage probability is 0.5. The intervals for $p = 0.5$ are considerably narrower than for $p = 0.05$.

The table suggests that widths are not greatly reduced unless reliability is rather high. The width is not halved until $\rho^4 = 0.75$, so that $\rho^2 = 0.866$. Even for a ρ^2 of 0.6, the width is 80% of

Table 3.
Widths of* 100(1-p)*% Prediction Intervals for Parallel Form Score and True Score for Standard Deviation of 100

	Form score		True score	
Reliability	$p = 0.05$	$p = 0.5$	$p = 0.05$	$p = 0.5$
0.0	392	135	392	135
0.1	390	134	372	128
0.2	384	132	351	121
0.3	374	129	328	113
0.4	359	124	304	104
0.5	339	117	277	95
0.6	314	108	248	85
0.7	280	96	215	74
0.8	235	81	175	60
0.9	171	59	124	43

the width for $\rho^2 = 0$.

Prediction intervals for the true score T_k that are based on the observed score X_{1k} are a bit smaller. Under the normal approximation, an interval that contains T_k with probability $1 - p$ has center $\mu + \rho(X_1 - \mu)$ and width $2z\sigma(1 - \rho^2)^{1/2}$. Results can be found in Table 3. Here, relative to the interval width for $\rho^2 = 0$, the interval width is halved if $\rho^2 = 0.75$, and the interval width is divided by 3 if $\rho^2 = 0.89$.

4. Conclusions

The proposed interpretations of reliability can be presented to indicate the consequences of reliability coefficients of various values to provide a test user a notion of reasonable expectations. On the whole, the measures presented appear to suggest relatively high standards for reliability coefficients, although different individuals may interpret the numerical results in quite distinct ways.

References

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, *53*, 814–861.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.